



## 2D-QSAR study of the antimicrobial activity of a series of 5-(substituted benzaldehyde) thiazolidine-2,4-dione derivatives against *Staphylococcus aureus* by Multiple Linear Regression method

Y. El Masaoudy<sup>1</sup>, I. Aanouz<sup>1</sup>, Y. Moukhliiss<sup>1</sup>, Y. Koubi<sup>1</sup>, H. Maghat<sup>1\*</sup>, T. Lakhlifi<sup>1</sup>  
and M. Bouachrine<sup>1,2</sup>

<sup>1</sup>Molecular Chemistry and Natural Substances Laboratory (MCNSL), Department of Chemistry, Faculty of Science,  
University of Moulay Ismail, Meknes, Morocco.

<sup>2</sup>Higher School of Technology (EST-Khenifra), University of Sultan Moulay Slimane, Benimellal, Morocco.

\*Corresponding Author: E-mail: [h.maghat@umi.ac.ma](mailto:h.maghat@umi.ac.ma)

Received 22 Sept 2020,

Revised 28 Nov 2020,

Accepted 01 Dec 2020

### Keywords

- ✓ 2D-QSAR,
- ✓ Antimicrobial activity,
- ✓ *S. aureus*,
- ✓ PCA,
- ✓ MLR.

[h.maghat@umi.ac.ma](mailto:h.maghat@umi.ac.ma)

Phone: +212662201441

### Abstract

In this research, the antimicrobial activity of 5-( substituted benzaldehyde) thiazolidine-2,4-dione derivatives against *Staphylococcus Aureus* has been submitted to a Two-Dimensional Quantitative Structure-Activity Relationship analysis. This analysis has been performed with two statistical methods integrated in the XLSTAT software. These methods are the Principal Component Analysis (PCA) and the Multiple Linear Regression (MLR). The 14 molecular descriptors involved in this study have been calculated using Gaussian 09, MarvinSketch and ACD/ChemSketch software programs. In order to select the molecular descriptors that exhibit a strong interrelationship with the experimental activity and no correlation between them, the Principal Component Analysis has been applied. The Multiple Linear Regression (MLR) has been employed to make correlation between antimicrobial activity pMIC with selected molecular descriptors of the studied compounds. The dataset of 20 compounds was arbitrarily divided into a training set (of 16 compounds), which has been employed to generate the QSAR model, and a test set (of 4 compounds) that has been used to assess the external predictive ability of the QSAR model. In this study, numerous statistical coefficients were used to select the best model ( $R=0.981$ ,  $R^2 = 0.963$ ,  $R^2_{adj} = 0.953$ ,  $MSE = 5.97 \times 10^{-5}$  and  $R^2_{test} = 0.997$ ). The QSAR model proposed via the MLR was validated internally and externally by several criteria, namely Y-randomization test,  $r_m^2$  and  $\Delta r_m^2$  metrics and Golbraikh-Tropsha's criteria. The domain of applicability of the proposed model has been applied using the Williams' diagram to identify the compounds that are outside this domain. The established QSAR model can be utilized to help predict antimicrobial activity of the studied molecules.

## 1. Introduction

*Staphylococcus aureus*, known as “Golden *Staphylococcus*”, is a Gram-positive bacterium responsible for the contagion of so many types of infections in the human body. It is among the most commonly isolated pathogens of hospital and community infections [1].

*Staphylococci* are commensal bacteria that colonize the nasal passages, vagina, pharynx, or damaged skin surfaces. Infections are triggered when a rupture of the skin or mucous barrier allows *staphylococci* to access adjacent tissues or blood circulation [2-3].

*S. aureus* is perhaps one of the greatest concerns due to its intrinsic virulence. It is able to cause a wide range of life-threatening infections, and it can adapt to different environmental conditions [4]. This bacteria pathogen is the main cause of infections in the blood circulation [5-6], pneumonia, skin and soft-tissue infection [7].

Various factors interpret the severity of *S. aureus* infections such as the ubiquitous character of the bacterium and the multi-resistance of certain strains to antibiotics. Though the pathogenicity of *S. aureus* is associated to the expression of virulence factors, after invasion, it has the capacity to secrete adhesion factors, toxins or enzymes. Toxic compounds can be classified as super antigens, or "pore forming toxins". Among these toxins, some are responsible for specific syndromes. These toxins have the ability to destroy host defense cells by forming pores at cell membranes. *S. aureus* secretes many toxins to divert or neutralize the immune response of the infected host [1-2-8].

The treatment of *S. aureus* infections has become more difficult. This creates therapeutic problems mainly due to its resistance to antibiotics [3]. Currently, there are different research studies that are working on the development of new molecules characterized by antimicrobial activity against the *S. aureus* such as a series of 21 5- (substituted benzaldehyde) thiazolidine-2,4-dione derivatives that were synthesized by Sucheta et al [9].

The objective of this analysis is to establish and evaluate the 2D-QSAR model for predicting antimicrobial activity of 5- (substituted benzaldehyde) thiazolidine-2,4-dione derivatives against *Staphylococcus aureus*. The model is developed using statistical methods, precisely the Principal Component Analysis (PCA) and the Multiple Linear Regression (MLR), and several molecular descriptors. The stability and the predictive power of the selected QSAR model are assessed using internal and external validation by Y-randomization test,  $r_m^2$  and  $\Delta r_m^2$  metrics [20-21], parameter  $R_p^2$  [18] and by Golbraikh-Tropsha's criteria [22- 23].

## 2. Material and Methods

### 2.1. Data sources

To build a Quantitative Structure – Activity Relationship, we choose 20 derivatives of 5- (substituted benzaldehyde) thiazolidine-2,4-dione exhibiting antimicrobial activity (MIC) against *S. aureus* reported by Sucheta et al [9], where (MIC) means the minimum concentration ( $\mu\text{M} / \text{mL}$ ) that blocks the growth of the bacterial strain *Staphylococcus aureus*. Figure 1 represents the general structure of the compounds whereas the structure of each compound and associated experimental activity (MIC) are presented in Table 1. The studied series consists of 21 compounds with MIC values of experimental activity ranging from  $4.2 \mu\text{M} / \text{mL}$  to  $5.65 \mu\text{M} / \text{mL}$ , except that compound N°12 displays high MIC value ( $10.6 \mu\text{M} / \text{mL}$ ), which is excluded from the series as an outlier.

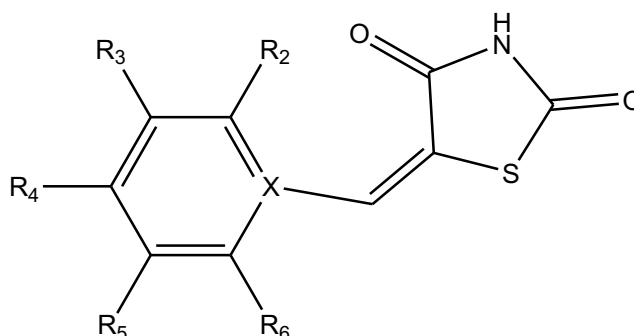



Figure 1: General chemical structure of the studied compounds.

**Table 1:** Antimicrobial activity values of the compounds.

N°	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	X	MIC(= μM/mL)
1	H	H	NO <sub>2</sub>	H	H	-	4.90
2	H	H	Cl	H	H	-	5.21
3	Cl	H	Cl	H	H	-	4.50
4	NO <sub>2</sub>	H	H	H	H	-	4.99
5	H	H	OH	H	H	-	5.65
6	H	H	N(CH <sub>3</sub> ) <sub>2</sub>	H	H	-	5.00
7	H	NO <sub>2</sub>	H	H	H	-	4.99
8	H	H	N(C <sub>2</sub> H <sub>5</sub> ) <sub>2</sub>	H	H	-	4.50
9	H	H	Br	H	H	-	4.30
10	H	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	-	4.20
11	H	OC <sub>2</sub> H <sub>5</sub>	OH	H	H	-	4.70
12	H	H	OCH <sub>3</sub>	H	H	-	10.60
13	Cl	H	H	H	H	-	5.21
14	H	Cl	H	H	H	-	5.21
15	OCH <sub>3</sub>	H	H	H	H	-	5.31
16	H	OCH <sub>3</sub>	H	H	H	-	5.31
17	OH	H	H	H	H	-	5.65
18	H	OCH <sub>3</sub>	OH	H	H	-	4.90
19	H	OCH <sub>3</sub>	OCH <sub>3</sub>	H	H	-	4.70
20	H	H	H	H	H		5.40
21	H	H	CHO	H	H	-	5.30

## 2.2. Calculation of molecular descriptors

In this stage of our study, we calculated 14 molecular descriptors for 20 compounds to establish the 2D-QSAR model using ChemSketch [10], MarvinSketch [11] and Gaussian 09 [12] software programs. The molecular descriptors used in the QSAR analysis are exhibited in Table 2. The calculation of electronic descriptors was carried out with the Gaussian 09 package. The geometric optimizations of the 20 compounds were performed by the Density Functional Theory (DFT) using Becke's three-parameter hybrid functional (B3LYP) together with 6-31G (d) basis set [13].

Additionally, four other quantum chemical descriptors: Energy Gap ( $\Delta E$ ), Hardness ( $\eta$ ), Electronegativity ( $\chi$ ) and Electrophilicity index ( $\omega$ ) [14] were calculated by the following formulas:

$$\Delta E = E_{LUMO} - E_{HOMO}$$

$$\eta = \frac{E_{LUMO} - E_{HOMO}}{2}$$

$$\chi = -\frac{(E_{LUMO} + E_{HOMO})}{2}$$

$$\omega = \frac{\chi^2}{2\eta}$$

**Table 2:** The molecular descriptors used in the QSAR study.

Software	Molecular descriptors
ChemSketch	Parachor (Pc), Polarizability (P), Density (d), Surface tension (S), and Index of refraction (n).
MarvinSketch	Partition coefficient (log P)
Gaussian 09	Energy Gap ( $\Delta E$ ), Total energy ( $E_T$ ), Lowest unoccupied molecular orbital energy ( $E_{LUMO}$ ), Highest occupied molecular orbital energy ( $E_{HOMO}$ ), Dipole moment ( $\mu$ ), Electrophilicity index ( $\omega$ ), Hardness ( $\eta$ ) and Electronegativity ( $\chi$ ).

### 2.3. Statistical Methods

The statistical data analysis methods are necessary to build the 2D-QSAR model between the dependent variable (biological activity) and the independent variables (molecular descriptors). In this research, we applied Principal Component Analysis (PCA) and Multiple Linear Regression (MLR) implemented in the XLSTAT software [15].

The Principal Component Analysis (PCA) is a tool used to sum up all the information encoded in the structures of molecules and also to understand the distribution of these molecules [16]. In this work, it can be used for examining the correlations between the molecular descriptors and the experimental activity in order to select the best molecular descriptors to be involved in the QSAR model development. The multiple linear regression (MLR) is the most frequently applied tool in Two-Dimensional Quantitative Structure-Activity Relationship due to its simplicity and easy interpretability. This method is used to link between several independent variables  $X_n$  and a dependent variable Y, according to the following mathematical form [17]:

$$Y = a_0 + \sum a_n X_n$$

Where  $Y$  is the experimental activity pMIC,  $a_0$  is the constant of the model and  $a_n$  are the coefficients of the descriptors  $X_n$ .

### 2.4. Methods of validation

In general, the statistical parameters of the best selected QSAR model, internal and external validation must be performed to check its quality, stability and predictive ability. First, in order to test the quality and the reliability of the selected model, different statistical parameters are used: the squared correlation coefficient  $R^2$ , the correlation coefficient R, the adjusted squared correlation coefficient  $R_{adj}^2$ , the mean squared error MSE and the coefficient of Fischer F. They are calculated according to following formulas:

$$R^2 = 1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{obs})^2}$$

$$R = \sqrt{1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - \bar{Y}_{obs})^2}}$$

$$R_{adj}^2 = \frac{(N - 1)R^2 - p}{N - p - 1}$$

$$MSE = \frac{\sum(Y_{obs} - Y_{pred})^2}{N}$$

$$F = \left(\frac{N - p - 1}{p}\right) \cdot \frac{\sum(Y_{pred} - \bar{Y}_{pred})^2}{\sum(Y_{obs} - Y_{pred})^2}$$

With:

$Y_{obs}$  and  $Y_{pred}$  are observed and predicted activity value, respectively.

$\bar{Y}_{obs}$  and  $\bar{Y}_{pred}$  are the average of the observed and predicted activity value, respectively.

$p$  represents the number of descriptors used in the model development and  $N$  represents the number of molecules.

Second, the internal predictive capacity of the model is tested on the set used to develop it (the training set) using the cross-validation (Leave-one out) by calculating the squared correlation coefficient  $Q^2$  according to the following expression:

$$Q^2 = 1 - \frac{\sum(Y_{pred(train)} - Y_{obs(train)})^2}{\sum(Y_{obs(train)} - \bar{Y}_{obs(train)})^2}$$

Where  $Y_{obs(train)}$  and  $Y_{pred(train)}$  indicate observed and predicted activity value respectively, and  $\bar{Y}_{obs(train)}$  is the average of the observed activity value [18].

Currently, the Y-randomization is a widely employed test to perform the internal validation of the QSAR model. This test aims to randomly shuffle the antimicrobial activity values of the molecules in the training set for building random models using the same molecular descriptors of the original model. The coefficients ( $R^2$ ,  $Q^2$ ) of the original model should be better than those of the new models ( $R_r^2$ ,  $Q_r^2$ ) for several random trials. In this case, the original model is considered robust [19]. We also used the parameter  $R_p^2$  which ensures if the model has been obtained by chance or not. This parameter is defined as follow:

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$$

Where  $R^2$  and  $R_r^2$  are the squared correlation coefficient of the non-randomized model and the average correlation coefficient of randomized models respectively [18].

A valid value of  $Q^2$  that meet the requirement of  $Q^2 > 0.5$  may not mean that the predicted activity data are close to the observed activity ones despite the existence of a high correlation among the values. To avoid this problem and to show the predictability model better, the metrics  $\bar{r}_m^2$  and  $\Delta r_m^2$  introduced by Roy et al [20-21] as shown in the equations bellow have been used:

$$\bar{r}_m^2 = \frac{|r_m^2 + r_m'^2|}{2}$$
$$\Delta r_m^2 = |r_m^2 - r_m'^2|$$

Where:  $r_m^2 = R^2 \times (1 - \sqrt{(R^2 - R_0^2)})$  and  $r_m'^2 = R^2 \times (1 - \sqrt{(R^2 - R_0'^2)})$

$R^2$  and  $R_0^2$  are the squared correlation coefficient between the observed and predicted values of the molecules with and without intercept, respectively.

$R_0'^2$  bears the same meaning, but uses the reversed axes.

$\bar{r}_m^2$  is the average value of  $r_m^2$  and  $r_m'^2$ .

$\Delta r_m^2$  is the absolute difference between  $r_m^2$  and  $r_m'^2$ .

The metrics  $\bar{r}_{m(train)}^2$  and  $\Delta r_{m(train)}^2$  are used in the case of internal validation.

The metrics  $\bar{r}_{m(test)}^2$  and  $\Delta r_{m(test)}^2$  are used in the case of external validation.

Third, according to Golbraikh and Tropsha [22-23], the developed model is regarded satisfactory and predictive once the following criteria are met:

$$Q^2 > 0.5 \text{ and } R_{test}^2 > 0.6$$
$$0.85 \leq k' \leq 1.15 \text{ and } 0.85 \leq k \leq 1.15$$
$$\frac{(R^2 - R_0^2)}{R^2} < 0.1$$

$$\frac{(R^2 - R_0'^2)}{R^2} < 0.1$$

$$|R_0^2 - R_0'^2| < 0.3$$

where  $Q^2$  is calculated for the training set, but  $R_{test}^2$ ,  $k$ ,  $k'$ ,  $R_0^2$  and  $R_0'^2$  are calculated for the test set as follows:

$$R_{test}^2 = 1 - \frac{\sum(Y_{pred(test)} - Y_{obs(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{obs(train)})^2}$$

$$k = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum Y_{pred}^2}$$

$$k' = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum Y_{obs}^2}$$

$$R_0^2 = 1 - \frac{\sum(Y_{pred} - k \times Y_{pred})^2}{\sum(Y_{pred} - \bar{Y}_{pred})^2}$$

$$R_0'^2 = 1 - \frac{\sum(Y_{obs} - k' \times Y_{obs})^2}{\sum(Y_{obs} - \bar{Y}_{obs})^2}$$

$k$  and  $k'$  are the slopes of the regression lines through the origin.

### 3. Results and discussion

#### 3.1. Dataset and descriptors

The antimicrobial activity ( $MIC$ ) values were converted into  $pMIC$  in ( $M/L$ ) ( $pMIC = -\log_{10} MIC$ ) which were employed for the QSAR study using 14 molecular descriptors (Table 3).

The dataset comprising 20 compounds was randomly divided into two sets: a training set consisting of 16 molecules was employed to establish the QSAR model, and a test set consisting of 4 molecules was employed to test the predictive ability of the proposed model.

Table 3: Values of experimental activity  $pMIC$  and descriptors calculated.

$N^\circ$	$pMIC$	$Pc$	$n$	$S$	$d$	$P$	$\log P$	$E_T$	$E_{HOMO}$	$\mu$	$E_{LUMO}$	$\Delta E$	$\eta$	$\chi$	$\omega$
1	2.310	466.30	1.732	78.10	1.595	24.9	1.79	-1193.19	-0.251	2.842	-0.119	0.132	0.066	0.185	0.259
2	2.283	446.40	1.707	65.40	1.527	24.2	1.79	-1448.29	-0.234	1.524	-0.095	0.139	0.070	0.165	0.195
3	2.347	483.50	1.710	67.10	1.622	26.2	3.06	-1907.89	-0.242	1.198	-0.097	0.145	0.073	0.170	0.198
4	2.302	466.30	1.732	78.10	1.595	24.9	1.79	-1193.19	-0.249	4.544	-0.107	0.142	0.071	0.178	0.223
5	2.248	424.40	1.744	76.70	1.542	23	1.55	-1063.92	-0.217	4.007	-0.082	0.135	0.068	0.150	0.166
6	2.301	513.90	1.697	62.20	1.356	28	1.96	-1122.68	-0.194	7.637	-0.071	0.123	0.062	0.133	0.143
7	2.302	466.30	1.732	78.10	1.595	24.9	1.79	-1193.19	-0.247	4.492	-0.105	0.142	0.071	0.176	0.218
8	2.347	594.00	1.660	57.20	1.279	31.6	2.67	-1201.03	-0.196	8.583	-0.078	0.118	0.059	0.137	0.159
9	2.367	460.30	1.724	66.50	1.762	25.3	2.62	-3559.80	-0.233	1.631	-0.095	0.138	0.069	0.164	0.195
10	2.377	585.10	1.620	52.80	1.360	30.2	1.38	-1332.26	-0.211	2.655	-0.085	0.126	0.063	0.148	0.174
11	2.328	523.10	1.679	65.40	1.442	27.5	1.75	-1217.77	-0.209	3.528	-0.082	0.127	0.064	0.146	0.167
13	2.283	446.40	1.707	65.40	1.527	24.2	2.45	-1448.29	-0.239	3.040	-0.090	0.149	0.075	0.165	0.182
14	2.283	446.40	1.707	65.40	1.527	24.2	2.45	-1448.29	-0.238	3.014	-0.096	0.142	0.071	0.167	0.196
15	2.275	467.80	1.667	58.70	1.391	24.9	1.69	-1103.23	-0.219	5.010	-0.079	0.140	0.070	0.149	0.159
16	2.275	467.80	1.667	58.70	1.391	24.9	1.69	-1103.22	-0.223	2.363	-0.089	0.134	0.067	0.156	0.182
17	2.248	424.40	1.744	76.70	1.542	23	1.55	-1063.92	-0.228	1.804	-0.088	0.140	0.070	0.158	0.178
18	2.310	483.10	1.701	69.30	1.500	25.7	1.39	-1178.45	-0.210	3.502	-0.082	0.128	0.064	0.146	0.167
19	2.328	526.50	1.640	55.30	1.374	27.6	1.53	-1217.74	-0.213	3.394	-0.084	0.129	0.065	0.149	0.171
20	2.268	477.00	1.761	72.80	1.416	26.7	2.38	-1066.11	-0.216	4.010	-0.092	0.124	0.062	0.154	0.191
21	2.276	456.40	1.737	72.00	1.488	25	1.56	-1102.02	-0.242	2.570	-0.108	0.134	0.067	0.175	0.229

### 3.2. Principal Component Analysis (PCA)

In this step, we studied the correlations between the 14 molecular descriptors in order to select the best among them to be included in the development of the QSAR model using the Principal Component Analysis tool. The correlation circle in Figure 2 and the correlation matrix in Table 4 examine the correlations between the molecular descriptors.

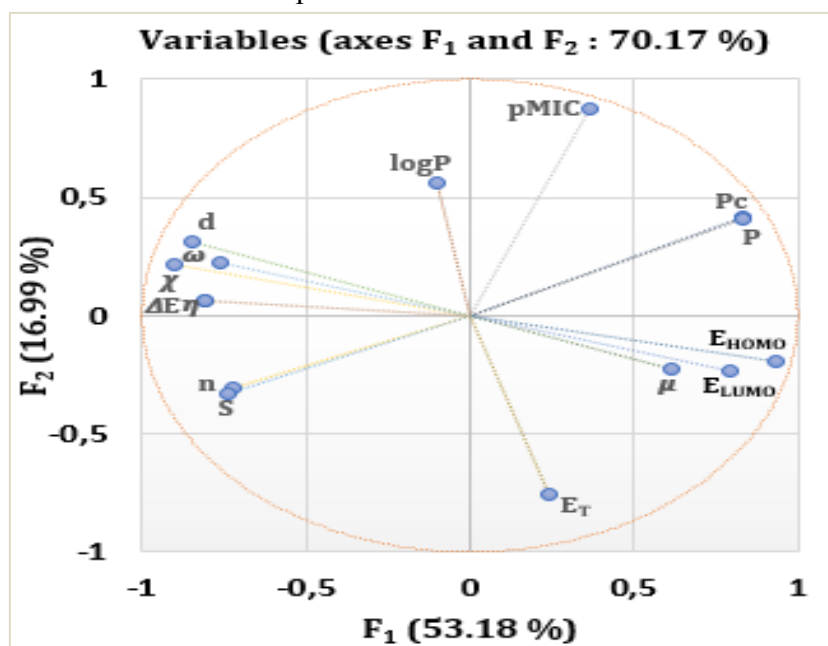


Figure 2: Correlation circle of the descriptors and antimicrobial activity.

Table 4: Correlation matrix of molecular descriptors.

Var	<i>pMIC</i>	<i>n</i>	<i>S</i>	<i>d</i>	<i>P</i>	<i>logP</i>	<i>Pc</i>	<i>E<sub>T</sub></i>	<i>E<sub>HOMO</sub></i>	<i>μ</i>	<i>E<sub>LUMO</sub></i>	<i>ΔE</i>	<i>η</i>	<i>χ</i>	<i>ω</i>
<i>pMIC</i>	1														
<i>n</i>	-0.539	1													
<i>S</i>	-0.466	<b>0.912</b>	1												
<i>d</i>	0.016	0.633	0.646	1											
<i>P</i>	0.706	-0.654	-0.638	-0.639	1										
<i>logP</i>	0.291	0.167	-0.066	0.246	0.160	1									
<i>Pc</i>	0.726	-0.741	-0.663	-0.647	<b>0.984</b>	0.048	1								
<i>E<sub>T</sub></i>	-0.531	-0.071	0.082	-0.623	<b>0.033</b>	-0.523	0.081	1							
<i>E<sub>HOMO</sub></i>	0.169	-0.522	-0.590	-0.755	0.645	-0.138	0.624	0.214	1						
<i>μ</i>	0.034	-0.193	-0.182	-0.586	0.534	0.058	0.494	0.367	0.577	1					
<i>E<sub>LUMO</sub></i>	0.044	-0.512	-0.609	-0.624	0.413	-0.067	0.406	0.132	<b>0.902</b>	0.456	1				
<i>ΔE</i>	-0.288	0.351	0.354	0.675	-0.750	0.191	-0.715	-0.257	-0.787	-0.547	-0.444	1			
<i>η</i>	-0.288	0.351	0.354	0.675	-0.750	0.191	-0.715	-0.257	-0.787	-0.547	-0.444	<b>1.000</b>	1		
<i>χ</i>	-0.120	0.530	0.613	0.719	-0.564	0.112	-0.548	-0.185	<b>-0.984</b>	-0.541	<b>-0.965</b>	0.663	0.663	1	
<i>ω</i>	-0.033	0.502	0.605	0.587	-0.375	0.046	-0.369	-0.105	-0.869	-0.419	<b>-0.997</b>	0.379	0.379	<b>0.943</b>	1

On the one hand, the circle shows that the  $F_1$  axis is the first dimension of the PCA and the  $F_2$  axis is the second dimension of the PCA representing 53.18% and 16.99% respectively of the total variance,



and both represent **70.17%** of the total information. The circle also shows three types of angles among the descriptors: an acute angle reflects a positive correlation between the molecular descriptors, a right angle separates two uncorrelated descriptors and an obtuse angle represents a negative correlation.

On the other hand, the matrix displays information on the strong and weak correlation between the molecular descriptors. If two descriptors are highly correlated ( $R > |\pm 0.9|$ ), the one which provides the greatest information is retained. According to the correlation coefficients, it turns out that the high correlation is observed between the Hardness ( $\eta$ ) and the Energy Gap ( $\Delta E$ ) ( $R = 1.000$ ), and the low correlation is observed between the total energy ( $E_T$ ) and the polarizability ( $P$ ) ( $R = 0.033$ ).

Based on the results of the matrix and the correlation circle, we selected three molecular descriptors that exhibit a high interrelationship with the experimental activity  $pMIC$  and no correlation between them: Parachor ( $Pc$ ), Total energy ( $E_T$ ) and Highest occupied molecular orbital energy ( $E_{HOMO}$ ).

### 3.3. Multiple Linear Regression (MLR)

The best QSAR model obtained by the Multiple Linear Regression links the selected molecular descriptors  $Pc$ ,  $E_T$  and  $E_{HOMO}$  to the antimicrobial activity  $pMIC$  of the molecules as shown in the following mathematical formula:

$$pMIC = 1.706 + 7.536 \times 10^{-4} \times Pc - 4.873 \times 10^{-5} \times E_T - 0.740 \times E_{HOMO}$$

The high values of  $R$ ,  $R^2$ ,  $R^2_{adj}$  and  $F$ , and the low value of  $MSE$  confirm that this model is statistically significant, so we can conclude that it is robust and possesses good quality. [Table 5](#) gives the values of the statistical parameters of the best QSAR model using the RLM.

**Table 5:** Values of statistical parameters obtained by the RLM model.

Statistical parameter	Value
$N_{train}$	16
$R$	0.981
$R^2$	0.963
$R^2_{adj}$	0.953
$MSE$	$5.97 \times 10^{-5}$
$F$	103.004

The multicollinearity among the molecular descriptors Parachor  $Pc$ , Total energy  $E_T$  and Highest occupied molecular orbital energy  $E_{HOMO}$  is verified by the variance inflation factor VIF. The VIF values of these descriptors model are less than **2.000** ( $VIF < 2$ ), which indicates that there is no collinearity among them. The VIF value of  $Pc$ ,  $E_T$  and  $E_{HOMO}$  involved in MLR model is shown in [Table 6](#). The importance of the molecular descriptors  $Pc$ ,  $E_T$  and  $E_{HOMO}$  included in this QSAR model on antimicrobial activity  $pMIC$  is assessed by the absolute value of the t-test as displayed in [Table 6](#). From the t-test values, it is clear that the molecular descriptors contribute to the explanation of antimicrobial activity as to the following order:

$$Pc > E_T > E_{HOMO}$$

**Table 6:** VIF and t-test value of descriptors involved in MLR model.

Descriptors	$Pc$	$E_T$	$E_{HOMO}$
t-test	13.964	-4.848	-4.748
VIF	1.758	1.275	1.959



The linear relationship between the observed  $pMIC_{(obs)}$  and predicted  $pMIC_{(pred)}$  antimicrobial activities of the two sets (the test set is represented in red and the training set in blue) established by RLM is shown in Figure 3.

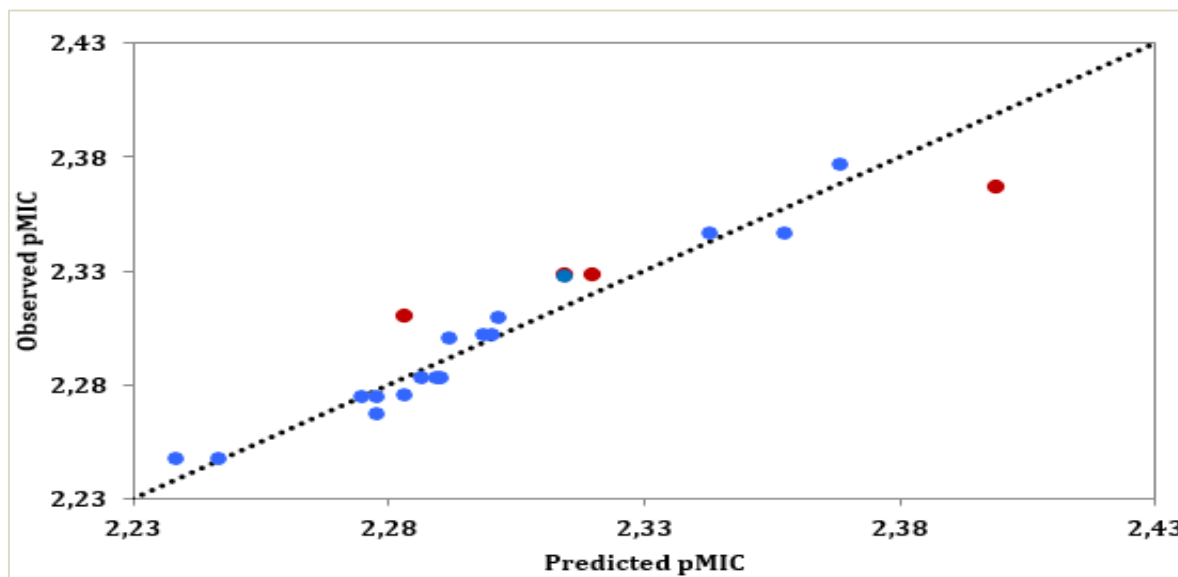


Figure 3: Graphical representation of observed and predicted activities.

The observed and predicted activity  $pMIC$  values for the training set and test set compounds obtained by the Multiple Linear Regression tool are presented in Table 7.

Table 7: Values of observed and predicted antimicrobial activity  $pMIC$ .

	N°	$pMIC_{(obs)}$	$pMIC_{(pred)}$	Residual
Training set	1	2.310	2.302	0.008
	2	2.283	2.286	-0.003
	3	2.347	2.343	0.004
	4	2.302	2.300	0.002
	5	2.248	2.238	0.009
	6	2.301	2.292	0.009
	7	2.302	2.299	0.003
	8	2.347	2.357	-0.011
	10	2.377	2.368	0.009
	13	2.283	2.290	-0.007
	14	2.283	2.289	-0.006
	15	2.275	2.275	0.000
	16	2.275	2.278	-0.003
	17	2.248	2.247	0.001
Test set	20	2.268	2.277	-0.010
	21	2.276	2.283	-0.007
	9	2.367	2.399	-0.032
	11	2.328	2.314	0.013
	18	2.310	2.283	0.027
	19	2.328	2.320	0.008

### 3.4. Y-randomization test

We applied a Y-randomization test on a set of 16 compounds to ensure the robustness of the QSAR model. The  $R_r^2$  and  $Q_r^2$  values for 50 random trials are lower than those ( $R^2$  and  $Q^2$ ) of the model proposed by the MLR, and the average values of  $R_r^2$  and  $Q_r^2$  of 50 randomized models are 0.170 and -0.575 respectively. Hence, the established QSAR model is not due to chance and is considered robust. The values of the statistical parameters  $R_r^2$  and  $Q_r^2$  of the randomized models obtained by the Y-randomization test are given in Table 8.

**Table 8:** Results of randomized models.

Random model	$R_r^2$	$Q_r^2$	Random model	$R_r^2$	$Q_r^2$
Rand 1	0.139	-0.394	Rand 26	0.115	-0.615
Rand 2	0.133	-1.166	Rand 27	0.216	-0.159
Rand 3	0.286	-0.353	Rand 28	0.023	-0.758
Rand 4	0.329	-0.025	Rand 29	0.270	-0.195
Rand 5	0.281	-0.435	Rand 30	0.075	-0.738
Rand 6	0.023	-0.465	Rand 31	0.232	-0.508
Rand 7	0.055	-0.905	Rand 32	0.079	-0.786
Rand 8	0.183	-0.782	Rand 33	0.227	-0.389
Rand 9	0.253	-0.387	Rand 34	0.003	-0.590
Rand 10	0.046	-0.973	Rand 35	0.206	-0.436
Rand 11	0.116	-0.381	Rand 36	0.006	-0.919
Rand 12	0.225	-0.518	Rand 37	0.221	-0.656
Rand 13	0.039	-0.681	Rand 38	0.160	-0.510
Rand 14	0.040	-0.575	Rand 39	0.243	-0.125
Rand 15	0.012	-0.806	Rand 40	0.063	-0.865
Rand 16	0.320	-0.739	Rand 41	0.047	-0.771
Rand 17	0.235	-0.336	Rand 42	0.335	-0.978
Rand 18	0.130	-0.457	Rand 43	0.023	-0.880
Rand 19	0.240	-0.393	Rand 44	0.083	-0.471
Rand 20	0.240	-0.455	Rand 45	0.235	-0.325
Rand 21	0.200	-0.758	Rand 46	0.130	-1.047
Rand 22	0.642	0.276	Rand 47	0.414	-0.124
Rand 23	0.058	-0.939	Rand 48	0.068	-1.607
Rand 24	0.252	-0.294	Rand 49	0.161	-0.917
Rand 25	0.257	-0.159	Rand 50	0.122	-0.274
		$R^2$		$Q^2$	
Average		0.170		-0.575	
Original		0.963		0.916	

### 3.5. Model validation results

The best QSAR model obtained was internally validated by the squared correlation coefficient  $Q^2$ , Y-randomization parameters ( $R_r^2$  and  $Q_r^2$ ), parameter  $R_p^2$ , and  $\bar{r}_m^2(\text{train})$  and  $\Delta r_m^2(\text{train})$  metrics, and externally validated by Golbraikh–Tropsha criteria and  $\bar{r}_m^2(\text{test})$  and  $\Delta r_m^2(\text{test})$  metrics. The threshold values for various statistical criteria of external and internal validation for the QSAR model are satisfied. The results obtained are shown in Table 9 indicate that the QSAR model proposed via the MLR method has good predictive ability.

**Table 9:** Internal and external validation results.

	Parameter	Value	Threshold
Internal validation	$Q^2$	0.916	$> 0.5$
	$R_r^2$	0.170	$< R^2$
	$Q_r^2$	-0.575	$< Q^2$
	$R_p^2$	0.846	$> 0.5$
	$\bar{r}_{m(train)}^2$	0.720	$> 0.5$
	$\Delta r_{m(train)}^2$	0.003	$< 0.2$
External validation	$R_{test}^2$	0.997	$> 0.5$
	$\bar{r}_{m(test)}^2$	0.882	$> 0.5$
	$\Delta r_{m(test)}^2$	0.110	$< 0.2$
	$ R_0^2 - R_0'^2 $	0.026	$< 0.3$
	$\frac{(R^2 - R_0^2)}{R^2}$	0.004	$< 0.1$
	$\frac{(R^2 - R_0'^2)}{R^2}$	0.029	$< 0.1$
	$k$	1.002	$0.85 \leq k \leq 1.15$
	$k'$	0.998	$0.85 \leq k' \leq 1.15$

### 3.6. Applicability domain

The last stage of our research is the domain of applicability; it was used to visualize all dataset molecules that exist inside and outside this area using the Williams' diagram, which represents the variations of the standardized residual ( $\pm\sigma$ ) as a function of the leverage  $h_i$  [24]. The prediction is considered reliable for each molecule that is within the domain of applicability. For this reason, the leverages values of all molecules in the data set are calculated as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, 2, \dots, k)$$

Where  $x_i$  represents the descriptor row vector of the compound,  $X$  is the  $k * p$  matrix of  $p$  model descriptor values for  $k$  training set compounds and the superscript  $T$  refers to the transpose of the matrix/vector [25-26].

The warning leverage value  $h^*$  was defined as:  $3p/k$ , with  $p$  and  $k$  are the number of model descriptors plus one and the number of training set compounds, respectively. The criteria for a response outlier in the QSAR model are given as follows: standardized residual greater than three standard deviation units  $> 3\sigma$  and  $h_i > h^*$  [27].

The diagram of Williams for the training set and the test set of the model QSAR is presented in Figure 4. In this graph the leverage threshold is 0.75 and the standardized residual is  $\pm 3$ . The standardized residual value of each training and test compound is between  $+3$  and  $-3$ . Additionally, the leverage value of all compounds in the training set does not reach the warning value. In contrast, the leverage value of compound N°9 in the test set exceeds the warning value. Therefore, this compound falls outside the QSAR AD as outlier.

Finally, the results obtained by the domain of applicability confirm the reliability of the QSAR model to predict the activity of the compounds with very high confidence.

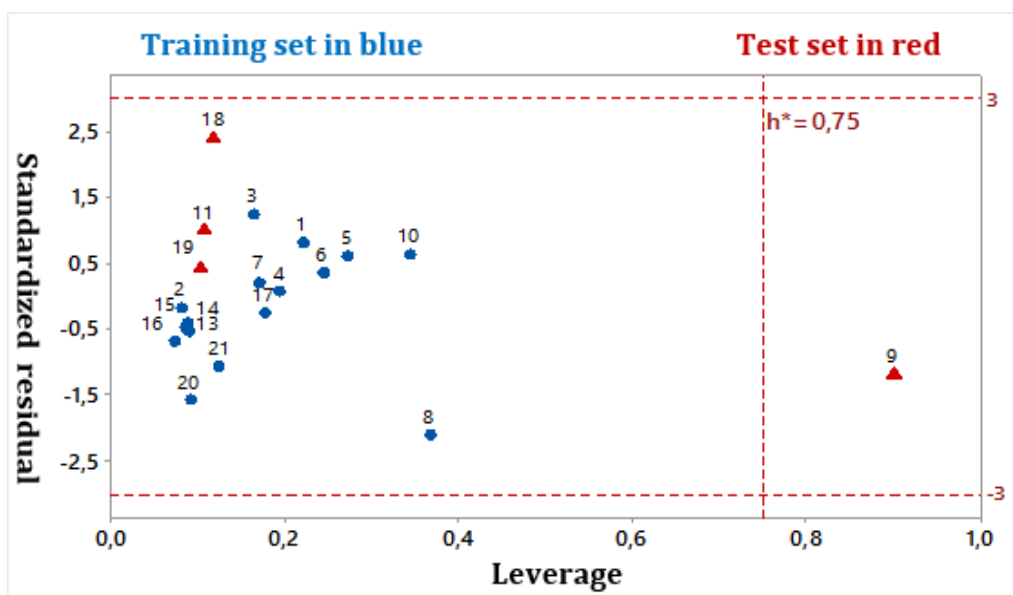


Figure 4: Williams' diagram for the MLR model (warning leverage  $h^* = 0.75$  ; residual limit  $\pm 3$ ).

## Conclusion

The 2D-QSAR study was used to construct a predictive model for the antimicrobial activity of a series of 5- (substituted benzaldehyde) thiazolidine-2,4-dione derivatives against *S. aureus*, using the selected descriptors: Highest occupied molecular orbital energy ( $E_{HOMO}$ ), Parachor (Pc) and Total energy ( $E_T$ ), which showed a strong interrelationship with the studied activity and there was no correlation between them. The internal and external validation criteria of the established model by the Multiple Linear Regression method were satisfied: ( $Q^2$ ,  $R_{test}^2$ ,  $R_p^2$ ,  $\bar{r}_{m(train)}^2$  and  $\bar{r}_{m(test)}^2$ )  $> 0.5$ , ( $\bar{r}_{m(train)}^2$  and  $\bar{r}_{m(test)}^2$ )  $< 0.2$ , Y-randomization parameters ( $R_r^2 < R^2$  and  $Q_r^2 < Q^2$ ) and Golbraikh-Tropsha's parameters. Finally, the results obtained from the internal validation, the external validation and the domain of applicability affirm that the QSAR model proposed via the MLR method is very promising as it can be utilized to predict antimicrobial activity of the studied compounds.

## References

1. O. Dumitrescu, O. Dauwalder, Y. Gillet, F. Vandenesch, J. Etienne, G. Lina, A. Tristan, "Les infections communautaires à *Staphylococcus aureus* en pédiatrie : émergence des staphylocoques dorés résistants à la méticilline d'origine communautaire," *Rev. Francoph. Lab.*, 407 (2008) 71–80, [https://doi: 10.1016/S1773-035X\(08\)74869-X](https://doi: 10.1016/S1773-035X(08)74869-X)
2. F. Vincenot, M. Saleh, G. Prévost, "Les facteurs de virulence de *Staphylococcus aureus*," *Rev. Francoph. Lab.*, 407(2008) 61–69, [https://doi: 10.1016/S1773-035X\(08\)74868-8](https://doi: 10.1016/S1773-035X(08)74868-8)
3. F. D. Lowy, "Staphylococcus aureus Infections," *N. Engl. J. Med.*, 339 (1998) 520–532, <https://doi: 10.1056/NEJM199808203390806>
4. S. Stefani, A. Goglio, "Methicillin-resistant *Staphylococcus aureus*: related infections and antibiotic resistance," *Int. J. Infect. Dis.*, 14 (2010) S19–S22, <https://doi: 10.1016/j.ijid.2010.05.009>
5. G. V. Doern, R. N. Jones, M. A. Pfaller, K. C. Kugler, M. L. Beach, and The SENTRY Study Group (North America), "Bacterial pathogens isolated from patients with skin and soft tissue infections: frequency of occurrence and antimicrobial susceptibility patterns from the SENTRY Antimicrobial Surveillance Program (United States and Canada, 1997)," *Diagn. Microbiol. Infect. Dis.*, 34 (1999) 65–72, [https://doi: 10.1016/S0732-8893\(98\)00162-X](https://doi: 10.1016/S0732-8893(98)00162-X)

6. D. J. Diekema, M.A. Pfaller, R.N. Jones, G.V. Doern, K.C. Kugler, M.L. Beach, H.S. Sader, "Trends in antimicrobial susceptibility of bacterial pathogens isolated from patients with bloodstream infections in the USA, Canada and Latin America," *Int. J. Antimicrob. Agents*, 13 (2000) 257–271, [https://doi: 10.1016/S0924-8579\(99\)00131-4](https://doi.org/10.1016/S0924-8579(99)00131-4)
7. M. A. Pfaller, R. N. Jones, G. V. Doern, K. Kugler, and T. S. P. Group, "Bacterial Pathogens Isolated from Patients with Bloodstream Infection: Frequencies of Occurrence and Antimicrobial Susceptibility Patterns from the SENTRY Antimicrobial Surveillance Program (United States and Canada, 1997)," *Antimicrob. Agents Chemother.*, 42 (1998) 1762–1770, [https://doi: 10.1128/AAC.42.7.1762](https://doi.org/10.1128/AAC.42.7.1762)
8. M. W. Parker, S. C. Feil, "Pore-forming protein toxins: from structure to function," *Prog. Biophys. Mol. Biol.*, 88(2005) 91–142, [https://doi: 10.1016/j.pbiomolbio.2004.01.009](https://doi.org/10.1016/j.pbiomolbio.2004.01.009)
9. Sucheta, S. Tahlan, and P. K. Verma, "Synthesis, SAR and in vitro therapeutic potentials of thiazolidine-2,4-diones," *Chem. Cent. J.*, 12 (2018) 129, [https://doi: 10.1186/s13065-018-0496-0](https://doi.org/10.1186/s13065-018-0496-0)
10. "Chemical Structure Drawing Software | ACD/ChemSketch." <https://www.acdlabs.com>
11. "ChemAxon - Software Solutions and Services for Chemistry & Biology." <https://chemaxon.com>
12. M. Frisch et al., "gaussian 09, Revision d. 01, Gaussian," Inc Wallingford CT, 2009.
13. E. Eroglu, H. Türkmen, "A DFT-based quantum theoretic QSAR study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme, CA-II," *J. Mol. Graph. Model.*, 26(2007) 701–708, [https://doi: 10.1016/j.jmglm.2007.03.015](https://doi.org/10.1016/j.jmglm.2007.03.015)
14. R.G. Parr, L.v. Szentpály, Shubin Liu, "Electrophilicity index," *J. Am. Chem. Soc.*, 121 (1999) 1922–1924, [https://doi: 10.1021/ja983494x](https://doi.org/10.1021/ja983494x)
15. "XLSTAT | Statistical Software for Excel," XLSTAT, Your data analysis solution. <https://www.xlstat.com>
16. M. Larif, A. Adad, R. Hmammouchi, A.I. Taghki, A. Soulaymani, A. Elmidaoui, M. Bouachrine, T. Lakhlifi, "Biological activities of triazine derivatives. Combining DFT and QSAR results," *Arab. J. Chem.*, 10 (2017) S946–S955, [https://doi: 10.1016/j.arabjc.2012.12.033](https://doi.org/10.1016/j.arabjc.2012.12.033)
17. T. K. Shameera Ahamed, V. K. Rajan, K. Muraleedharan, "QSAR modeling of benzoquinone derivatives as 5-lipoxygenase inhibitors," *Food Sci. Hum. Wellness*, 8 (2019) 53–62, [https://doi: 10.1016/j.fshw.2019.02.001](https://doi.org/10.1016/j.fshw.2019.02.001)
18. P. Pratim Roy, S. Paul, I. Mitra, K. Roy, "On Two Novel Parameters for Validation of Predictive QSAR Models," *Molecules*, 14 (2009) 1660–1701, [https://doi: 10.3390/molecules14051660](https://doi.org/10.3390/molecules14051660)
19. Y. Liu, Z. Ke, J. Cui, W.-H. Chen, L. Ma, B. Wang, "Synthesis, inhibitory activities, and QSAR study of xanthone derivatives as  $\alpha$ -glucosidase inhibitors," *Bioorg. Med. Chem.*, 16(2008) 7185–7192, [https://doi: 10.1016/j.bmc.2008.06.043](https://doi.org/10.1016/j.bmc.2008.06.043)
20. K. Roy, S. Kar, R. N. Das, "Statistical Methods in QSAR/QSPR," in *A Primer on QSAR/QSPR Modeling*, Springer International Publishing, (2015) 37–59, [https://doi: 10.1007/978-3-319-17281-1\\_2](https://doi.org/10.1007/978-3-319-17281-1_2)
21. K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, S. Kar, R. N. Das, "Some case studies on application of 'rm<sup>2</sup>' metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data," *J. Comput. Chem.*, 34 (2013) 1071–1082, [https://doi: 10.1002/jcc.23231](https://doi.org/10.1002/jcc.23231)
22. A. Golbraikh, A. Tropsha, "Beware of q<sup>2</sup>!," *J. Mol. Graph. Model.*, 20 (2002) 269–276, [https://doi: 10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
23. A. Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation," *Mol. Inform.*, 29 (2010) 476–488, [https://doi: 10.1002/minf.201000061](https://doi.org/10.1002/minf.201000061)

24. S. Chtita, A. Aouidate, A. Belhassan, A. Ousaa, A.I. Taourati, B. Elidrissi, M. Ghamali, M. Bouachrine, T. Lakhlifi, “QSAR study of N-substituted oseltamivir derivatives as potent avian influenza virus H5N1 inhibitors using quantum chemical descriptors and statistical methods,” *New J. Chem.*, 44 (2020) 1747–1760, <https://doi: 10.1039/C9NJ04909F>
25. T. I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, Weida Tong, G. Veith, C. Yang, “Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships” *Altern. Lab. Anim.*, 33 (2005) 155–173, <https://doi: 10.1177/026119290503300209>
26. P. Gramatica, “Principles of QSAR models validation: internal and external,” *QSAR Comb Sci*, 26 (2007) 694–701, <https://doi: 10.1002/qsar.200610151>
27. L. Qin, X. Zhang, Y. Chen, L. Mo, H. Zeng, Y. Liang, “Predictive QSAR Models for the Toxicity of Disinfection Byproducts,” *Molecules*, 22 (2017) 1671, <https://doi: 10.3390/molecules22101671>

(2020) ; <http://www.jmaterenvirosci.com>